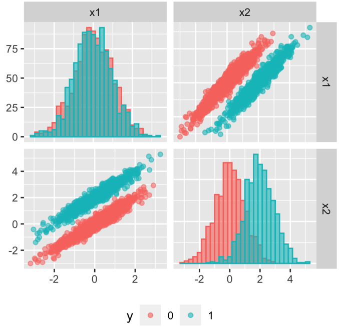
Supervised Learning

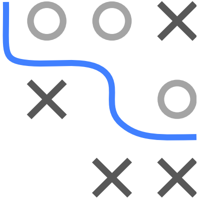
Filter Methods: Examples and Caveats



Learning goals

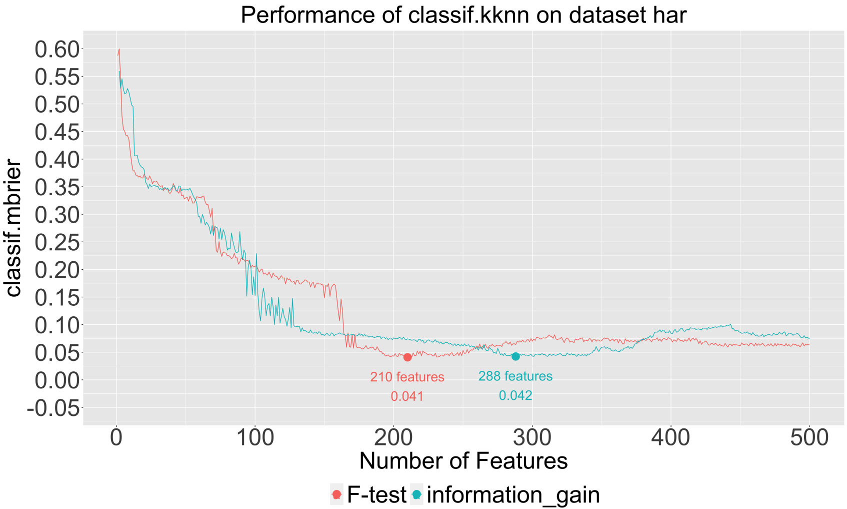
- Understand how filter methods can be misleading.
- Understand how filter methods work in practical applications.

INTRODUCTION

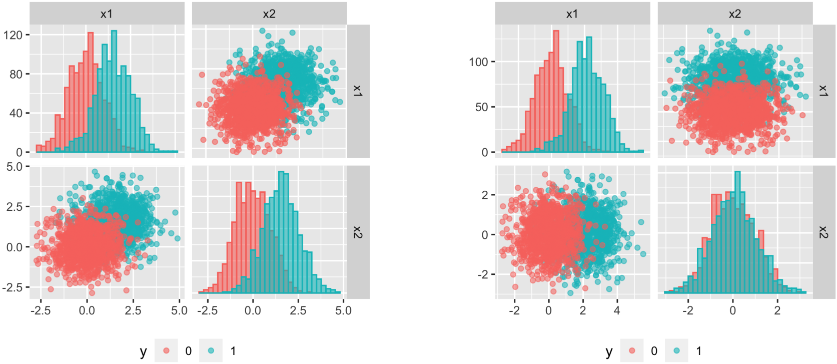


- **Filter methods** construct a measure that quantifies the dependency between all features and the target variable.
- They yield a numerical score for each feature x_j, according to which we rank the features.
- They are model-agnostic and can be applied generically.
- Filter methods are strongly related to methods for determining variable importance.

VISUALIZATION OF FILTER ALGORITHMS

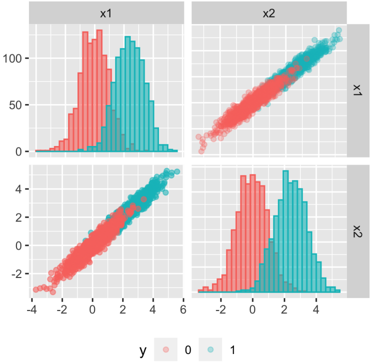


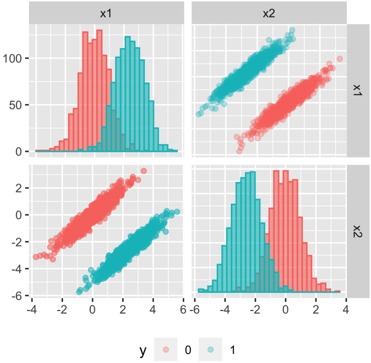
FILTER METHODS CAN BE MISLEADING



IG from presumably redundant variables. Left: 2 class problem with i.i.d. variables. Each class has Gaussian distr. with no covariance. Right: After 45 degree rotation, showing combination of 2 vars yields separation improvement by factor $\sqrt{2}$, showing i.i.d. vars are not truly redundant. For further details, see ▸ Guyon and Elisseeff, 2003.

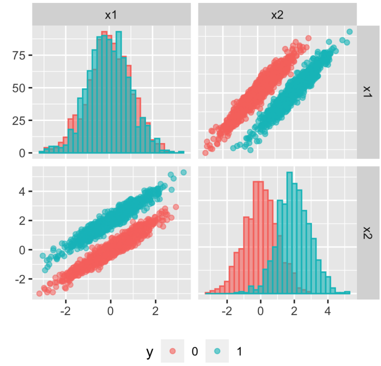
FILTER METHODS CAN BE MISLEADING

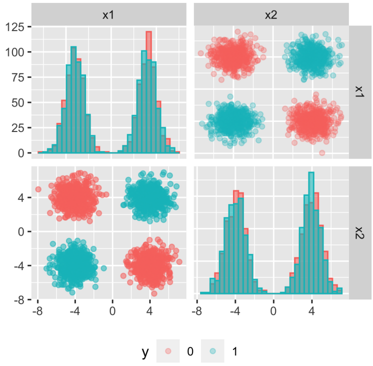


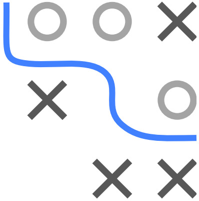


Intra-class covariance. In projection onto the axes, distribution of two variables are same as before. Left: Class conditional distribution have high cov. in direction of the line of the two class centers. Right: Class conditional distr. have high cov. in direction perpendicular to line of two class centers. Important separation gain is obtained by using both variables.

FILTER METHODS CAN BE MISLEADING







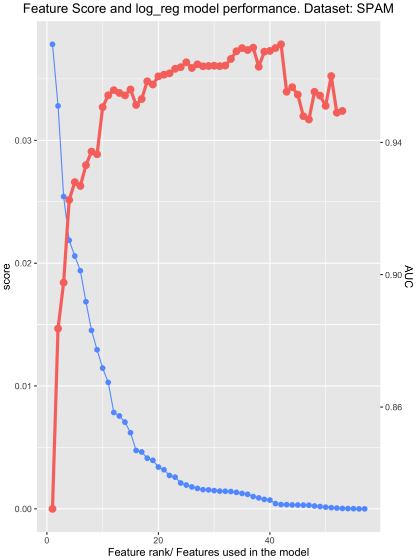
Variable useless by itself can be useful together with others. Left: One var has completely overlapping class conditional densities. Still, jointly with other variable separability can be improved. Right: XOR-like chessboard problem. Classes consist of “clumps" s.t. projection on the axes yield overlapping densities. Single vars have no separation power, only used together.

USING FILTER METHODS

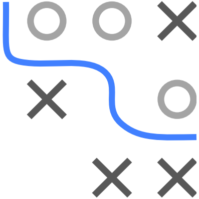
1. Calculate filter score for each feature x_j.
2. Rank features according to score values.
3. Choose $\tilde{p}$ best features.
4. Train model on $\tilde{p}$ best features.

How to choose $\tilde{p}$?

- It can be prescribed by the application.
- Eyeball estimation: read from filter plots
- Use resampling.



USING FILTER METHODS



Advantages:

- Easy to calculate.
- Typically scales well with the number of features *p*.
- Generally interpretable.
- Model-agnostic.

Disadvantages:

- Univariate analyses may ignore multivariate dependencies.
- Redundant features will have similar weights.
- Ignores the learning algorithm.